README and Guidance

Andrew Foote, Mark Kutzbach, Lars Vilhuber

2020-09-24

Contents

This README describes the data inputs and processing stream for our paper "*Recalculating . . . : How Uncertainty in Local Labor Market Definitions Affects Empirical Findings*".

Data Availability and Provenance Statements

Commuting Zone Data

- Source: Economic Research Service (2012) (https://www.ers.usda.gov/data-products/commuting-zones-and-labor-market-areas/)
- Source URL: https://www.ers.usda.gov/webdocs/DataFiles/48457/czlma903.xls?v=6997.1
- **Provided** as part of this replication package.
- Datafile: `czlma903.xls`

CZ data were produced by an agency of the US Government and are in the public domain.

Journey-to-Work (JTW) data

Most of the JTW data can be found at https://www.census.gov/topics/employment/commuting/guidance/flows.html. The data were produced by an agency of the US Government and are in the public domain.

Because the US Census Bureau does not provide robust (permanent) URLs, we archived the data on openICPSR/DataLumos, or searched for permanent locations elsewhere on ICPSR. As of 2020-09-01, the source URLs were still functional, though. Our scripts pull the data from the source URL.

1990 JTW

- Source: U.S. Census Bureau (2017a)
- Source URL: https://www2.census.gov/programs-surveys/commuting/datasets/1990/worker-flow/usresco.txt
- Permanent Source URL: http://doi.org/10.3886/E100617V1
- Not provided as part of this replication package
- Renamed to: `1990jtw_raw.txt`

2000 JTW

- Source: U.S. Census Bureau (2003)

- Source URL: https://www.census.gov/population/www/cen2000/commuting/files/2KRESCO_US.txt
- Permanent Source URL: http://doi.org/10.3886/ICPSR13405.v1
- Not provided as part of this replication package
- Renamed to : `jtw2000_raw.txt`

2009-2013 ACS flows

- Source: U.S. Census Bureau (2017b)
- Source URL: https://www2.census.gov/programs-surveys/commuting/tables/time-series/commuting-flows/table1.xlsx
- Permanent Source URL: http://doi.org/10.3886/E100616V1
- Renamed to: `jtw2009_2013.csv`
- Not provided as part of this replication package

Files for Case Study 1

BEA data Data on National Income and Product Accounts (NIPA). Used in replications.

- Source: Bureau of Economic Analysis (2019)
- Source URL: https://apps.bea.gov/regional/zip/CAINC30.zip.
  - Note: Data can be downloaded from https://apps.bea.gov/regional/downloadzip.cfm, under "Personal Income (State and Local)", select CAINC30: Economic Profile by County, then download. A direct download is also possible, see next line. The file is regularly updated.
- The datafile is **provided** as part of this package.
- Datafile: `CAINC30__ALL_AREAS_1969_2018.csv`

The data were produced by an agency of the US Government and are in the public domain.

BLS Data (Quarterly Census of Employment and Wages)

Data from Quarterly Census of Employment and Wages (QCEW) program

- Source: Bureau of Labor Statistics (2020)
- Source URL: https://www.bls.gov/cew/downloadable-data-files.htm
- Note: Data are downloaded using programs provided in Vilhuber and Bjelland (2020) (not part of this archive), see https://github.com/labordynamicsinstitute/readin_qcew_sas/releases/tag/v20200622 (also https://doi.org/10.5281/zenodo.3903458).
- The full data are not provided as part of this package.
  - Note: For convenience, the extract used is **provided** in `$interwrk` (`bls_us_county.dta.gz`), but must be unzipped prior to use. If using, the QCEW-related programs in Case Study 1 should not be run.

The data were produced by an agency of the US Government and are in the public domain.

ADH-related data files

- Note: We thank David Dorn for generously providing us with some of his data files.

NHGIS data

- Source: Minnesota Population Center (2016)
- Raw data are provided as part of this package, as per NHGIS permission to post extracts for the purpose of replication packages.
- Datafile: `$raw/nhgis/*.dta`

NIH/NCI SEER county population estimates

- Source: National Cancer Institute (2020)
- Original Source URL: https://seer.cancer.gov/popdata/yr1990_2018.singleages/us.1990_2018.singleages.adjusted.txt.gz
- Our Source URL: https://data.nber.org/seer-pop/uswbosingleagesadj.dta.zip
- Raw data is not provided as part of this package, but a derived file (`popcounts.dta`) is provided in $interwrk.
- Datafile: `popcounts.dta`

The data were produced by an agency of the US Government and are in the public domain.

1990 Counties to 1990 Commuting Zones

- Source: Dorn (n.d.)
- Source URL: https://www.ddorn.net/data/cw_cty_czone.zip
  - Note: Dorn references Autor and Dorn (2013b) for this file, which in turn has replication package Autor and Dorn (2013a). The replication package contains a file `cw_puma1990_czone.dta` which would seem to provide the same information. However, we downloaded directly from David Dorn's website Dorn (n.d.) , file [E7]
- The datafile is not provided as part of this package.
- Datafile: `cw_cty_czone.zip`

Before re-using this data, ask David Dorn for permission. Posted here with permission.

County-level industry data

- Source: Dorn (2017)
- Source URL: Email from David Dorn. See ddorn/README.md.
- The datafiles are provided as part of this package.
- Datafiles: `$raw/ddorn/cty_industryYYYY.dta`

Before using this data, ask David Dorn for permission. Posted here with permission.

China Syndrome Data

- Source: Autor, Dorn, and Hanson (2013b) and its replication package Autor, Dorn, and Hanson (2013a)
- Source URL: https://www.ddorn.net/data/Autor-Dorn-Hanson-ChinaSyndrome-FileArchive.zip
  - Note: the files are also archived at Autor, Dorn, and Hanson (2013a).
- The datafiles are NOT provided as part of this package.
- Datafiles: `$raw/adh_data/Public Release Data/dta/sic87dd_trade_data.dta` and `$raw/adh_data/Public Release Data/dta/workfile_china.dta`

Dataset list

The following files are provided in `$raw` directory:

filename
ddorn/cty_industry1980.dta
ddorn/cty_industry1990.dta
ddorn/cty_industry2000.dta
nhgis/nhgis0008_ds95_1970_county.dat
nhgis/nhgis0008_ds98_1970_county.dat
nhgis/nhgis0008_ds99_1970_county.dat
nhgis/nhgis0009_ds122_1990_county.dat
nhgis/nhgis0009_ds123_1990_county.dat
nhgis/nhgis0010_ds146_2000_county.dat

filename
nhgis/nhgis0010_ds151_2000_county.dat
nhgis/nhgis0011_ds195_20095_2009_county.dat
nhgis/nhgis0011_ds196_20095_2009_county.dat
nhgis/nhgis0012_ds103_1980_county.dat
nhgis/nhgis0012_ds107_1980_county.dat
CAINC30___ALL_AREAS_1969_2018.csv
czlma903.xls

The following files are provided in `$interwrk` directory. They can be recreated from files in `$raw` using various programs, and are provided as a convenience.

filename
07_adh_cutoff_post.dta
bartik_results_cutoff.dta
bartik_results_moe_new.dta
bls_us_county.dta.gz
bootstrap_results.dta
finalstats_jtw1990_moe_new2.dta
popcounts.dta

Data Created by this Archive

Commuting flows augmented by MOE

Filename: `flows_jtw1990_moe.{csv,dta,sas7bdat}`

Variables:

- `work_cty`: FIPS code of work county
- `jobsflow`: flows (count) between `work_cty` and `home_cty`
- `home_cty`: FIPS code of home county
- `flowsize`: categorical flow sizes (1: 0-9, 2: 10-136, 3: 137-454, 4: 455-6714, 5: 6715-max)
- `sd_ratio`:
- `mean_ratio`:
- `draw`:
- `moe`: Margin of error for flows as computed (see text)

Sample observations:

work_cty	jobsflow	home_cty	flowsize	sd_ratio	mean_ratio	draw	moe
31137	8	40097	1	0.48832	1.62034	2.12948	17.03581
25021	6	25023	1	0.48832	1.62034	1.76572	10.59431
23021	2	23021	1	0.48832	1.62034	0.77939	1.55878
26161	9	12095	1	0.48832	1.62034	1.26426	11.37833
23025	2	23021	1	0.48832	1.62034	2.04119	4.08237
20091	5	26161	1	0.48832	1.62034	1.50346	7.51730

Clusters for 1990 created by our algorithm

Filename: `clusfin_jtw1990.{csv,dta,sas7bdat}`

Variables:

- `_PARENT_` : Character cluster number (CL + NNNNN or CL + "10" + NNNNN)
- `_NAME_`: Character county FIPS code (cty + NNNNN)
- `county`: county FIPS code (numeric part, NNNNN)
- `cluster`: numeric cluster number (numeric part, NNNNN or "10" + NNNNN)

The naming convention for the commuting zones is CL + (fips of largest county by residence labor force). For singletons, the commuting zone is named CL + "10" + fips, to distinguish it from clusters in other realizations in which that county is the largest unit.

Sample observations:

PARENT	*NAME*	county	cluster
CL625	cty39007	39007	625
CL625	cty27143	27143	625
CL625	cty08017	08017	625
CL625	cty08061	08061	625
CL625	cty08011	08011	625
CL625	cty08099	08099	625

Bootstrap cluster assignments

This dataset contains the 1000 realizations of the commuting zones from our paper. It can be used to crosswalk county fips codes to commuting zone realizations.

Filename: `bootclusters_jtw1990_moe.{csv,sas7bdat}` (for technical reasons, the `dta` file has a `_new` suffix)

Variables:

- `fips` : county FIPS code (numeric part, NNNNN)
- `clustername`: character cluster number (CL + NNNNN)
- `clustername_Z`: character cluster number for *Z-th* draw (CL + NNNNN)

Software Requirements

- SAS 9.4 (TS1M0)
  - SAS/STAT 12.3 (maintenance)
- Stata 14.2/16.1
- R 4.0.2 (used only to automate cleaning of one data file)
  - readxl, tidyr, dplyr, readr for processing
  - rprojroot, config for configuration
  - all dependencies are installed upon first run
- Bash, Curl, wget as part of download (may require Linux, but can be replaced by manual downloading)

Memory and Runtime Requirements

These programs were last run as follows:

- OS: Linux CentOS release 6.3 (Final)
- 8-core (though probably only 1 core was in use)
- 147 GB RAM (unlikely to have been fully utilized)
- about 1.5GB disk space required

Description of programs

Setting up data

To create the commuting zone analysis, data download programs (and in some cases, cleaning programs) are in the `raw` folder. They are not downloaded by the SAS and Stata programs in the `$programs` folder. Download is accomplished using Linux tools, but can also be done by hand, using the URLs mentioned above or in the scripts.

filename
01_get_data.sh
02_convert.R
03_get_adh.sh
nhgis/main.sh
nhgis/nhgis0008_ds95_1970_county.do
nhgis/nhgis0008_ds98_1970_county.do
nhgis/nhgis0008_ds99_1970_county.do
nhgis/nhgis0009_ds122_1990_county.do
nhgis/nhgis0009_ds123_1990_county.do
nhgis/nhgis0010_ds146_2000_county.do
nhgis/nhgis0010_ds151_2000_county.do
nhgis/nhgis0011_ds195_20095_2009_county.do
nhgis/nhgis0011_ds196_20095_2009_county.do
nhgis/nhgis0012_ds103_1980_county.do
nhgis/nhgis0012_ds107_1980_county.do

Notes:

- QCEW: Data are downloaded using programs provided in Vilhuber and Bjelland (2020) (not part of this archive), see https://github.com/labordynamicsinstitute/readin_qcew_sas/releases/tag/v20200622 (also https://doi.org/10.5281/zenodo.3903458).
- NHGIS: See raw/nhgis/README.nhgis.txt for details
- ADH data: Files are downloaded and unpacked using `raw/03_get_adh.sh`. If processing manually, see URL above, and unzip into directory called `adh_data`. The resulting data structure should look like this:

```
$raw/adh_data/Public Release Data/dta
```

Main program files

The main program files are split into three groups: the creation and analysis of the commuting zones, for which all programs are in the main `$programs` directory, and case studies 1 (QCEW) and 2 (ADH). The programs for each of the case studies are in subdirectories `06_qcew` and `07_adh`, respectively.

In all cases, programs should be executed in the numeric sequence implied by the name of the program. If programs have the same numeric prefix, they can be executed in any order, or in parallel.

Setting up programs

- modify `config.sas`:
  - change the line with `root =` to correspond to your project directory
- modify `config.do`:
  - change the line with `root =` to correspond to your project directory

Order of programs to run

To create the replicated commuting zones, run the following programs in numerical order:

filename
01_dataprep.sas
02_01_clusters.sas
02_02_export_data.sas
03_prep_figures.sas
04_figures2_3.do
05_01_flows.do
05_02_bootstrap.sas
05_03_export_bootstraps.sas
05_04_bootstrap_graphs_new.do
08_map_inset.sas
09_maps_paper.sas
config.do
config.sas

Reading in various datasets

```
sas 01_dataprep.sas
```

(runtime: 2.81s)

Clustering process

```
sas 02_01_clusters.sas
```

(runtime: 3:25.73 minutes)

OUTPUT: $data/clusfin_jtw1990.sas7bdat

Outputting other formats

```
sas 02_02_export_data.sas
```

(runtime: 1.35s)

OUTPUT: $data/clusfin_jtw1990.{csv,dta}

Cutoff by Cluster Count (Figure)

```
sas 03_prep_figures.sas
```

(runtime: 8:39 minutes)

```
stata -b do 04_figures2_3.do
```

(runtime: seconds)

Run the Bootstrap Projects MOEs from 2009-2013 onto 1990 data, creates the 1000 realizations of commuting zones.

```
stata -b do 05_01_flows.do
sas         05_02_bootstrap.sas
```

The first program runs in seconds, the second one takes *(runtime: 56 hours)*.

Figure 4

`stata -b do 05_03_bootstrap_graphs_new.do`

(runtime: seconds)

Replication programs for Case Study 1 in Section 4.1

All programs are in `$programs/06_qcew/` subdirectory. Change working directory, and execute in numerical order.

Data preparation Required data are commuting zones, BEA-collected receipt of UI benefits (Bureau of Economic Analysis 2019), QCEW employment data (Bureau of Labor Statistics 2020).

Programs prefixed with `00` prepare the data:

filename
06_qcew/00_bea_readin.do
06_qcew/00_describe_bootclusters.do
06_qcew/00_qcew_extraction.sas
06_qcew/00_qcew_post_extraction.do
06_qcew/00_readin_czones.do

Analysis programs The remaining programs generate the analysis described in the manuscript, and output tables and figures as per the list below. Programs with non-numeric prefixes are called by other programs, and should not be run separately. Scripts (`*.sh`) are for convenience, and are not necessary - simply execute all programs in numerical order.

filename
06_qcew/01_regressions_table.do
06_qcew/02_01_cluster_loop.do
06_qcew/02_02_cluster_loop.do
06_qcew/03_01_cluster_graphs.do
06_qcew/03_02_cutoff_graphs.do
06_qcew/zz_bartik_merge.do

The complete sequence of programs ran in about 36 hours.

Replication programs for Case Study 2 in Section 4.2

All programs in `$programs/07_adh/` subdirectory. Change working directory, and execute in numerical order.

Data preparation Required data are commuting zones, and various ADH-related data listed earlier.

Programs prefixed with `00` prepare the data:

filename
07_adh/00_01_census_creation.do
07_adh/00_02_ctyindustry_creation.do
07_adh/00_03_IPW_creation.do
07_adh/00_04_cbp_readin.do
07_adh/00_05_subset_qcewdata.do
07_adh/00_06_subset_seerpop.do

filename
07_adh/00_07_mergecounty.do
07_adh/00_08_cz_merge.do

Analysis programs The remaining programs generate the analysis described in the manuscript, and output tables and figures as per the list below. Programs with non-numeric prefixes are called by other programs, and should not be run separately. Scripts (`*.sh`) are for convenience, and are not necessary - simply execute all programs in numerical order.

filename
07_adh/01_table3.do
07_adh/02_01_cutoff_loop.do
07_adh/02_02_overall_loop.do
07_adh/03_01_cutoff_graphs.do
07_adh/03_02_overall_graphs.do
07_adh/zz_aggregatedata.do
07_adh/zz_ctymerge.do

The complete sequence of programs ran in about 36 hours.

List of tables and programs

Figure/Table #	Title	Program	Output file
Figure 1 – left	Replication of Commuting Zones from TS: County Mapping	09_maps_paper.sas	ersamutingzones.png
Figure 1 – right	Replication of Commuting Zones from TS: County Mapping	02_clusters.sas	1990_replicationmap.png
Figure 2	Effect of Cluster Height on Number of Clusters	04_figures2_3.do	domclus_cutoff.pdf
Figure 3	Cluster Height and Share Workers Commuting Between Clusters	04_figures2_3.do	flows_cutoff.pdf
Figure 4	Results from Re-sampling Commuting Flows	05_03_bootstrap_graphs.do	mapchanges_jtw1990.pdf meanclussize_jtw1990.pdf mismatch_jtw1990.pdf
Figure 5	Differences in Effect Based on Cluster Cutoff	06_qcew/03_02_cutoff_graphs.do	cutoff_bartik.pdf
Figure 6	Distribution based on Realizations of CZs	06_qcew/03_01_cutoff_graphs.do	betacluster_distribution.pdf tdistribution_bartik.pdf
Figure 7	Differences in Effect Based on Cluster Cutoff	07_adh/03_01_cutoff_graphs.do	cutoff_iqr_1990.png

Figure/Table #	Title	Program	Output file
Figure 8	Distribution of Effect, 1990-2000	07_adh/03_02_1990_all_graphics.do	1990_all_distribution.png 1990_tstat_distribution.png
Table 1	Replication of TS1990 Commuting Zones: Summary Statistics	NA	NA
Table 2	Effect of Labor Demand on Unemployment Receipt	06_qcew/01_regressions_table.do	06_qcew/01_regressions_table.log
Table 3	China Syndrome Replication and Comparison, 1990-2000	07_adh/01_table3_adh.do	07_adh/01_table3.log
Figure A1	Clusters in California at Incremental Height Cutoffs	08_map_insets_cali.do	california_clustermap_800_inset6.png california_clustermap_880_inset6.png california_clustermap_1000_inset6.png california_clustermap_960_inset6.png
Figure A2	Hierarchical Clustering, Cutoff = 0.945	09_maps_paper_irwa1990_highcutoff.do	irwa1990_highcutoff
Table A1 (4)	Summary Statistics of Ratio of MOE to Flows	NA	NA
Table A2 (5)	Summary Statistics for empirical example	NA	NA

References

Autor, David H., and David Dorn. 2013a. "Replication Data for: The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market." American Economic Association [publisher]. https://doi.org/10.3886/E112652V1.

———. 2013b. "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market." *American Economic Review* 103 (5): 1553–97. https://doi.org/10.1257/aer.103.5.1553.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013a. "Replication Data for: The China Syndrome: Local Labor Market Effects of Import Competition in the United States." [Datafiles]. American Economic Association [publisher] ICPSR - Interuniversity Consortium for Political and Social Research [distributor]. https://www.openicpsr.org/openicpsr/project/112670/version/V1/view.

———. 2013b. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review* 103 (6): 2121–68. https://doi.org/10.1257/aer.103.6.2121.

Bureau of Economic Analysis. 2019. "Table 30: Economic Profile by County, 1969-2018." [Datafile]. U.S. Department of Commerce [producer]. https://apps.bea.gov/regional/zip/CAINC30.zip.

Bureau of Labor Statistics. 2020. "Quarterly Census of Employment and Wages – Data Files." [Datafiles]. Department of Labor [distributor]. https://www.bls.gov/cew/downloadable-data-files.htm.

Dorn, David. 2017. "County-Level Industry Data." [Dataset]. (provided via email).

———. n.d. "1990 Counties to 1990 Commuting Zones." [Datafile] [E7]. David Dorn's Data Page. Accessed September 20, 2020. https://www.ddorn.net/data.htm.

Economic Research Service. 2012. "1980 and 1990 Commuting Zones and Labor Market Areas." [Dataset]. United States Department of Agriculture. https://www.ers.usda.gov/webdocs/DataFiles/48457/czlma903.xls?v=7728.8.

Minnesota Population Center. 2016. "National Historical Geographic Information System." Minneapolis, MN: University of Minnesota. https://doi.org/10.18128/D050.V11.0.

National Cancer Institute. 2020. "U.S. Population Data (County-Level)- SEER Population Data." [Datafile] 1990-2018. National Bureau of Economic Research [distributor]. https://data.nber.org/seer-pop/.

U.S. Census Bureau. 2003. "Census of Population and Housing, 2000 [United States]: County-to-County Worker Flow Files: Version 1." [Datafile]. U.S. Department of Commerce [producer]. https://doi.org/10.3886/ICPSR13405.V1.

———. 2017a. "1990 County-to-County Worker Flow Files." [Datafile]. U.S. Department of Commerce [producer]. https://doi.org/10.3886/E100617V1.

———. 2017b. "2009-2013 5-Year American Community Survey: Commuting Flows." [Datafile]. U.S. Department of Commerce [producer]. https://doi.org/10.3886/E100616V1.

Vilhuber, Lars, and Melissa Bjelland. 2020. "Labordynamicsinstitute/Readin_qcew_sas: A Sequence of Programs to Readin in QCEW Data from the Bureau of Labor Statistics." Labor Dynamics Institute, Cornell University. https://doi.org/10.5281/zenodo.3903458.